FREL: A Stable Feature Selection Algorithm

Yun Li, *Member, IEEE*, Jennie Si, *Fellow, IEEE*, Guojing Zhou, Shasha Huang, and Songcan Chen

***Abstract*— Two factors characterize a good feature selection algorithm: its accuracy and stability. This paper aims at introducing a new approach to stable feature selection algorithms. The innovation of this paper centers on a class of stable feature selection algorithms called feature weighting as regularized energy-based learning (FREL). Stability properties of FREL using L1 or L2 regularization are investigated. In addition, as a commonly adopted implementation strategy for enhanced stability, an ensemble FREL is proposed. A stability bound for the ensemble FREL is also presented. Our experiments using open source real microarray data, which are challenging high dimensionality small sample size problems demonstrate that our proposed ensemble FREL is not only stable but also achieves better or comparable accuracy than some other popular stable feature weighting methods.**

***Index Terms*— Energy-based learning, ensemble, feature selection, feature weighting, uniform weighting stability.**

I. Introduction

FEATURE selection has been an active research area in machine learning and data mining for decades. It is an important and frequently used technique for data dimension reduction by removing irrelevant and redundant information from a data set. It is also a knowledge discovery tool for providing insights on the problem through interpretations of the most relevant features [1]. Discussions on feature selection usually center on two technical aspects: search strategy and evaluation criteria. Algorithms designed with different strategies broadly fall into three categories: filter, wrapper, and hybrid or embedded models [2]. On the other hand, if the categorization is based on output characteristics, feature selection algorithms can be divided into either feature weighting/ranking algorithms or subset selection algorithms. In this paper, we focus on feature weighting. A comprehensive survey of existing feature selection techniques and a general framework for their unification can be found in [1]–[3].

In addition to classification accuracy, another important measure is stability when evaluating the quality of a feature selection algorithm. Here, stability means the insensitivity of the result of a feature selection algorithm to variations in the training data set [4]. This issue is particularly important for some applications where feature selection is used as a knowledge discovery tool for identifying characteristic markers to explain the observed phenomena. A feature selection algorithm without stability constraint usually results in significantly different feature subsets due to variations in the training data. Even though most of these feature subsets are as good as they can be in terms of classification accuracy, unstable feature selection results can shake the confidence of domain experts when experimentally validating the selected features to interpret important discoveries [5]. For instance, in analyzing cancer biomarkers, such as leukemia, the available data sets usually are high dimensional yet with small sample size. Among the thousands of genetic expression levels, a critical subset is to be discovered that links to two leukemia labels. It is therefore necessary that the selected predictive genes are common to variations of training samples. Otherwise, the results will lead to less confident diagnosis. In consideration of the importance of stability in applications, several stable feature selection algorithms have been proposed. The ensemble methods [4], [6]–[8], sample weighting [9], [10], and feature grouping [5], [11] are a few examples. A comprehensive survey of earlier work can be found in [12]. Those existing stable feature selection algorithms make use of empirical criteria for stability measurements, and they fell short of explicitly providing a stability analysis. The pressing need for an analytical examination of stable feature selection algorithms beyond the simple empirical approach is thus evident.

In this paper, guided by energy-based learning [13], a new algorithm framework for feature weighting as regularized energy-based learning (FREL) is proposed. Stability of the proposed FREL algorithms under an L1 or L2 regularizer is examined. In addition, an ensemble FREL is also introduced and analyzed for its stability. The proposed FREL is then applied to open source real microarray data to demonstrate its effectiveness for both stability and accuracy in high dimensionality small sample size (HDSSS) application problems.

This paper is organized as follows. The framework of FREL and ensemble FREL are introduced in Section II. Section III analyzes the stability of feature weighting with an L1 or L2 regularizer. In addition, the stability analysis of ensemble FREL is presented. The experimental results on microarray data are shown in Section IV. This paper concludes in Section V.

II. Energy-Based Learning for Feature Weighting

Energy-based learning [13] provides a unified framework for many probabilistic and nonprobabilistic approaches to

Manuscript received August 20, 2013; revised April 21, 2014; accepted July 15, 2014. Date of publication August 12, 2014; date of current version June 16, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 60973097, Grant 61035003, Grant 61073114, Grant 61170151, and Grant 61300165, in part by the National Science Foundation of Jiangsu Province under Grant BK20131378 and Grant BK20140885, in part by the Jiangsu Government Scholarship, and in part by the Jiangsu Qinglan Project.

Y. Li, G. Zhou, and S. Huang are with the Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: liyun@njupt.edu.cn).

J. Si is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: si@asu.edu).

S. Chen is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: s.chen@nuaa.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2014.2341627

learning for prediction, classification, decision-making, sample ranking, detection, and conditional density estimation. In this paper, we consider an energy-based learning framework for the design of feature weighting algorithms. Specifically, we will focus on developing FREL. An ensemble FREL will also be discussed. In addition to these feature weighting algorithms as well as the implementations of these algorithms, this paper provides stability analysis for these algorithms.

A. Energy-Based Learning

Consider an inference model between input variable x and inferred variable y. The goodness of fit of each possible model configuration relating x to y can be measured by an energy function $E(y, x)$. The value of this energy function can be viewed as the degree of compatibility of a given configuration between x and y. Conventionally, small energy values correspond to highly compatible configurations, while large energy values correspond to highly incompatible configurations. When applying such an inference model, for a given input x, the model produces the most compatible answer $y^\star$ such that $y^\star = \text{argmin}_{y\in Y} E(y, x)$. The energy-based learning entails finding an energy function that produces the best y for a given x. To search for the best energy function, a family of parameterized energy functions of the form $\mathfrak{F} = \{E(w, y, x): w \in W\}$ is proposed, where w is the model parameter [13].

B. Regularized Energy-Based Learning

To train an energy-based model, we are given a training set D containing n samples, $D = \{X, Y\} = \{x_i, y_i\}_{i=1}^n$, where x_i is the ith training input and y_i is the corresponding desired answer such as a label, but not limited to that. Each sample input is represented by a d-dimensional vector, i.e., $x_i \in \mathbb{R}^d$. To find the best energy function in the family $\mathfrak{F}$, we need to assess the quality of an energy function, only with information from the training set D and possible prior knowledge about the task where data were collected. This quality measure is a loss functional, i.e., a function of functions, denoted by $L_D(w)$. We call it the objective loss function. Accordingly, the learning problem becomes finding the w' that minimizes the objective loss

$$w' = \text{argmin}_{w\in W} L_D(w). \tag{1}$$

Usually, an objective loss function based on data set D is defined as follows:

$$L_D(w) = \frac{1}{n}\sum_{i=1}^{n} L(w, x_i) + \gamma\,\mathcal{R}(w). \tag{2}$$

On the right-hand side of (2), $L(w, x_i)$ is the per-sample loss function. Then, the first term $(1/n)\sum_{i=1}^n L(w, x_i)$ is the sample-averaged loss function, which is taken over n respective per-sample loss function, and is denoted by $\mathcal{J}_D(w)$ for simplicity

$$\mathcal{J}_D(w) = \frac{1}{n}\sum_{i=1}^{n} L(w, x_i). \tag{3}$$

The $\mathcal{R}(w)$ in (2) is a regularizing term that can be used to embed prior knowledge about which energy functions are preferable to others. In this paper, the classical L1 and L2 regularizer are respectively examined. Parameter γ in (2) is a cost balancing factor.

Based on the discussion of energy-based learning above, it is evident that the per-sample loss function should be designed in such a way that it assigns a low loss to well-behaved energy functions, i.e., the energy functions that give the lowest energy to the correct answers and higher energy to all other including incorrect answers. Conversely, the energy functions that do not assign the lowest energy to the correct answers would have a high loss [13]. The generalized margin loss functions, for example, meet those conditions [13]. It is thus used as the per-sample loss function $L(w, x_i)$. Before introducing the generalized margin loss function, the following definition is needed.

Definition 1: Let y be a discrete variable. *The most offending incorrect answer* is the one that has the lowest energy among all the answers that are incorrect

$$\overline{y}_i = \text{argmin}_{y\in Y, y\neq y_i} E(w, y, x_i). \tag{4}$$

Then, a generalized margin loss function for per-sample loss can be described as follows:

$$L(w, x_i) = Q_\theta(E(w, y_i, x_i), E(w, \overline{y}_i, x_i)) \tag{5}$$

where $E(w, y_i, x_i)$, which is consequently denoted as E_i for notation simplification, is the energy of a correct answer for x_i; $E(w, \overline{y}_i, x_i)$, denoted by $\overline{E}_i$, is the energy of the most offending incorrect answer for x_i; θ is a positive parameter called the margin, and it is the energy gap between the incorrect answers and the correct ones. As discussed in [13], the function $Q_\theta(E_i, \overline{E}_i) \to \mathbb{R}$ is assumed to be convex, which can be easily satisfied. Moreover, consider the energy space defined by $E_i \times \overline{E}_i$, and let $\partial Q_\theta/\partial E_i$ and $\partial Q_\theta/\partial \overline{E}_i$ denote the gradient of Q_θ along E_i and $\overline{E}_i$, respectively. Then, in general, it holds true that $\partial Q_\theta/\partial E_i - \partial Q_\theta/\partial \overline{E}_i > 0$ in the region where $E_i + \theta > \overline{E}_i$ in the energy space. This means wherever $\overline{E}_i$ is smaller than E_i plus θ, the gradient along E_i is larger than the gradient along $\overline{E}_i$. Then, Q_θ pushes down the value of E_i and pulls up the value of $\overline{E}_i$. This causes the Q_θ loss surface to be slanted toward low values of E_i and high values of $\overline{E}_i$ [13]. This meets the specification in Section II-A that the energy value between x_i and its compatible answer y_i be small, while the energy value between x_i and its incompatible answer $\overline{y}_i$ is large.

Remark 1: There are many possible realizations of Q_θ in (5) such as the hinge, log, square-square, and square-exponential losses [13]. For example, when the log loss with infinite margin and the square-square loss with margin θ are selected, the corresponding per-sample loss functions in (5) are as follows:

$$L(w, x_i) = \log(1 + \exp(E(w, y_i, x_i) - E(w, \overline{y}_i, x_i))) \tag{6}$$

$$L(w, x_i) = E(w, y_i, x_i)^2 + (\max(0, \theta - E(w, \overline{y}_i, x_i)))^2. \tag{7}$$

If the parameter w in energy function E is defined as the feature weight vector, then a feature weighting algorithm simply finds the w' that minimizes the objective loss function defined in (2). In Section IV-D, the log and square-square losses are chosen as two representative per-sample loss

functions to construct the objective loss functions and to derive specific feature weighting algorithms. However, the theoretical results obtained in this paper are not limited to log and square-square losses in the feature weighting stability analysis.

C. Feature Weighting as Regularized Energy-Based Learning

In short, for a feature weighting problem to be considered an energy-based learning problem, first, the parameter w in the energy function should be relevant to the feature weight vectors. Second, a generalized margin loss function should be selected as per-sample loss function $L(w, x_i)$ as shown in (2) to construct the objective loss function $L_D(w)$. After that, the correct answer and the most offending incorrect answer for a sample should be explicitly identified. To summarize, the following issues are critical to the feature weighting algorithm design: 1) using appropriate criteria to determine the correct answer and the most offending incorrect answer for each sample and 2) properly designing the structure of an energy function E, which is needed in the per-sample loss function of (5) and consequently (2). By properly addressing these issues, learning leads to finding the appropriate energy function such that the per-sample loss function of (5) will be pushed down for correct answers and pulled up for incorrect answers while maintaining an energy gap or margin.

To address the first issue of developing appropriate criteria to determine the correct answer and the most offending incorrect answer for each sample, we resort to the nearest neighbor (NN) classification scheme. Note that, the NN classifier is a nonlinear mapping between input patterns and class labels. It is a simple algorithm but has received considerable attention again recently since they have been demonstrated highly efficient in some state-of-the-art real-world applications [14], [15]. Additionally, the NN classifier can be viewed as an energy-based learning where the energy function is a sample distance measure. Consider a sample x_i, its NN in the same class denoted by $N_{\mathcal{H}}(x_i)$ can be determined easily so long as a distance measure is defined. Also, the $N_{\mathcal{H}}(x_i)$ can be considered as the nearest correct answer. Similarly, the NN with a different label denoted by $N_{\mathcal{M}}(x_i)$, can be considered as the most offending incorrect answer based on Definition 1.

Once the correct and the most offending incorrect answers are identified as $N_{\mathcal{H}}(x_i)$ and $N_{\mathcal{M}}(x_i)$, respectively, parameterized energy functions $E(w, y_i, x_i)$ and $E(w, \overline{y}_i, x_i)$ needed in the generalized margin loss function of (5) can be defined as weighted Manhattan distances as shown in the following:

$$E(w, y_i, x_i) = E(w, N_{\mathcal{H}}(x_i), x_i) = w^T |x_i - N_{\mathcal{H}}(x_i)| \quad (8)$$

$$E(w, \overline{y}_i, x_i) = E(w, N_{\mathcal{M}}(x_i), x_i) = w^T |x_i - N_{\mathcal{M}}(x_i)| \quad (9)$$

where $|\cdot|$ denotes an element-wise absolute value operator on each component of the argument vector, which is of dimension d in the cases of (8) and (9), T denotes transpose, and w is the parameter associated with the energy function. As discussed above, the solution to the general energy-based learning problem with energy function shown in (2) becomes the feature weighting vector $w = \{w(1), w(2), \ldots, w(d)\}$ in our current problem setting.

Algorithm 1 FREL Algorithm

Step 1. Input training data set $D = \{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, margin θ in (5) and regularization parameter γ in (2).

Step 2. Initialize $w = (1, 1, \ldots, 1) \in \mathbb{R}^d$.

Step 3. For $i = 1, 2, \ldots, n$

(a) Given x_i, find the $N_{\mathcal{H}}(x_i)$ and $N_{\mathcal{M}}(x_i)$ based on NN algorithm.

(b) From (8) and (9), calculate $E(w, N_{\mathcal{H}}(x_i), x_i)$, $E(w, N_{\mathcal{M}}(x_i), x_i)$ and obtain the generalized margin loss Q_θ, i.e., $L(w, x_i)$, in (5).

(c) $\nabla = \frac{1}{n} \frac{\partial L(w, x_i)}{\partial w} + \gamma \frac{\partial \mathcal{R}(w)}{\partial w}$.

(d) $w = w - \frac{\nabla}{||\nabla||_2}$.

Step 4. Output the feature weighting vector $w' = w$.

The optimal feature weight w' can then be found by many different optimization approaches. As an example, the gradient descent algorithm is used to illustrate the minimization of the objective loss function (2) next. With the above discussions in place, we are in a position to summarize our proposed FREL as Algorithm 1.

Once the parameterized energy functions E in (8) and (9) are defined, respectively, we can employ a generalized margin loss function of the form (5) as the per-sample loss function $L(w, x_i)$. As discussed, many different generalized margin loss functions mentioned in Remark 1 such as the hinge and the log losses can be integrated with different regularizers, (e.g., L1 or L2 regularizer) to make up different objective loss functions $L_D(w)$ in (2). Therefore, a family of feature weighting algorithms could consequently be derived. Note that the local learning-based feature weighting algorithm described in [16] is a special case of FREL when the log loss and L1 regularizer are adopted. Moreover, if the log loss is combined with L2 regularizer, which is used to enhance diversity among base feature selectors in ensemble feature selection, the algorithm in [7] can be obtained and it is another special case of FREL. Note, however, this is the first time that the algorithms, including those in [7] and [16], are analyzed from an energy-based learning perspective under the proposed framework of FREL, and their respective stability properties are provided.

It also should be pointed out that for the purposes of this paper, we use Manhattan distance to determine the NNs and to define energy functions in (8) and (9). Nonetheless, other standard distance measures such as Euclidean distance are also eligible candidates without creating any problem in obtaining the results in this paper.

To summarize, resorting to NN classification, the feature weighting problem is described as regularized energy-based learning and the feature weighting vector corresponds to the parameter w in the objective loss function defined in (2). The generalized margin loss function Q_θ, which is convex, in (5) is adopted as per-sample loss function $L(w, x_i)$ in energy-based learning. For the regularizer in the objective loss function (2), the classical L1 and L2 regularizers are considered in this paper.

Algorithm 2 Ensemble FREL

Step 1. Input training data set $D=\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, margin θ in (5), regularization parameter γ in (2), random sampling parameters α and m.
Step 2. Initialize $w_{\mathcal{E}} = (0, 0, \ldots, 0) \in \mathbb{R}^d$.
Step 3. For $t = 1, 2, \ldots, m$
(a) Produce a bootstrap subset $D(r_t)$ with size $\lceil \alpha n \rceil$.
(b) Perform FREL on $D(r_t)$ to obtain a base weighting result $w_{D(r_t)}$.
(c) $w_{\mathcal{E}} = w_{\mathcal{E}} + \frac{1}{m} w_{D(r_t)}$.
Step 4. Output the ensemble feature weighting result $w_{\mathcal{E}}$.

D. Ensemble FREL

Ensemble learning is an effective approach for producing robust and accurate learning solutions in machine learning [17], [18] as demonstrated by many significant applications [19]–[21]. For instance, a popular ensemble learning making use of the bagging approach [22] consists in averaging several estimators built from random subsamples of the original data set.

Similar to the ensemble models for supervised learning, there are two essential steps in ensemble feature selection. The first step involves creating a set of different base feature selectors, each provides its output, while the second step aggregates the results of all base feature selectors [4].

In our case, bootstrap-based strategy as in [4] and [7] is used to train base feature selectors derived from FREL on m different bootstrap subsets of the original training set $D = \{x_i, y_i\}_{i=1}^n$. Ensemble feature weighting result is achieved by averaging the obtained solutions from the base feature selectors. Let $0 < \alpha < 1$ and $\lceil \alpha n \rceil$ be the integer closest to αn. For $t = 1, \ldots, m$, let $r_t = \{r_t(1), r_t(2), \ldots, r_t(\lceil \alpha n \rceil)\}$ be an index sequence randomly drawn from the natural sequence $\{1, \ldots, n\}$ without replacement. We denote the m bootstrap subsets by $D(r_t) = \{x_{r_t(k)}, y_{r_t(k)}\}_{k=1}^{\lceil \alpha n \rceil}$ for $t = 1, \ldots, m$ and the subsets are all drawn independently.

Let $w_{D(r_t)}$ denote the outcome of the feature weighting algorithm after FREL is applied on the tth bootstrap training subset $D(r_t)$. Therefore, we obtain m base feature weighting results $\{w_{D(r_1)}, w_{D(r_2)}, \ldots, w_{D(r_m)}\}$. In this paper, the ensemble result is obtained as

$$w_{\mathcal{E}} = \frac{1}{m} \sum_{t=1}^{m} w_{D(r_t)} \tag{10}$$

which is aggregated by averaging the outputs of base feature selectors. The pseudocode for the above discussed ensemble FREL is provided in Algorithm 2.

III. Stability Analysis

In this paper, the stability of FREL is considered in the following sense: variations in outputs are small or bounded in response to small variations in the input of the data set. This may entail the following two scenarios. The first is perturbation at the instance level caused by, for example, removing samples from or adding samples to the data set. The second is perturbation at the feature level caused by, for example, adding noise to the features in the data set. In addition, a combination of both types of perturbations may impose on a data set and cause stability concerns [4].

The stability of several classification, regression, and sample ranking methods has been analyzed thoroughly [23]–[25] in the sense similar to that described above. However, the stability of feature selection algorithms has only been examined empirically. This paper aims at providing a theoretical analysis for the stability of some feature weighting algorithms under FREL.

To account for small instance level perturbations, we only need to consider removing one sample from the data set and then analyzing the stability property of a feature weighting algorithm. Stability consideration after adding a sample follows directly from the result of removing a sample. To account for small feature level perturbations, we need to consider changing one sample and examine its impact on the stability of the algorithm. Before we proceed to analyzing both scenarios described above, consider the following.

For a given training set D of size n from a certain distribution $\mathbb{P}$, its samples are drawn independent identically distributed (i.i.d.) from $\mathbb{P}$. Let $D^{\setminus i}$ denote a modified training data set by removing the ith training sample (x_i, y_i) from the original training data set D, where $i \in \{1, \cdots, n\}$. We denote by D^i, the training set obtained by changing one sample from (x_i, y_i) to (x_i', y_i').

Definition 2: Consider a feature weighting algorithm A with output feature weight vectors denoted by w_D and $w_{D^{\setminus i}}$ for data set D and $D^{\setminus i}$, respectively. Algorithm A is *uniformly weighting stable* with stability bound β ($\beta \geq 0$) if for any D of size n and any $i \in \{1, \ldots, n\}$, we have

$$||w_D - w_{D^{\setminus i}}||_2 \leq \beta. \tag{11}$$

Intuitively, a smaller value of β corresponds to greater stability. To consider stability properties of algorithm A with the ith data sample distorted from the original ith data sample, i.e., the training data set is changed from D to D^i, we let the feature weight vector w_{D^i} denote the output of algorithm A for data set D^i. Based on the uniform weighting stability definition in (11) and by applying the triangle inequality, we have

$$\begin{aligned} ||w_D - w_{D^i}||_2 &= ||(w_D - w_{D^{\setminus i}}) - (w_{D^i} - w_{D^{\setminus i}})||_2 \\ &\leq ||w_D - w_{D^{\setminus i}}||_2 + ||w_{D^i} - w_{D^{\setminus i}}||_2 \\ &\leq 2\beta. \end{aligned} \tag{12}$$

Therefore, according to (12), the stability of changing one sample can be reduced for analyzing the stability of removing one sample. In other words, the uniform weighting stability formulated under the removal of one training data sample implies the same stability concept under the condition of one sample deviates from the original data sample. As such, stability in the sense of Definition 2 can be used for analyzing the stability of algorithm A under the perturbations at both instance and feature levels, which are described at the beginning of this section.

In the following sections, we will discuss the stability of FREL with L2 and L1 regularizers, and the stability of ensemble FREL, respectively.

A. Stability Analysis for FREL With L2 Regularizer

In this section, we examine stability properties of FREL with $\mathcal{R}(w)$ in (2) being an L2 regularizer.

Remark 2: For this part of the stability analysis, given the choice of the energy functions as in (8) and (9), we use the shorthand notation of $L(w^T z_i)$ in place of the per-sample loss function $L(w, x_i)$ in (5), where z_i is considered a transformation of x_i. For different loss functions in Remark 1, the expressions of z_i are different. For instance, if the log loss is used, then $z_i = |x_i - N_{\mathcal{H}}(x_i)| - |x_i - N_{\mathcal{M}}(x_i)|$. Furthermore, if the training samples x_i's are bounded and can be normalized, then $||z_i||_2$ should be as well. We denote this by $||z_i||_2 \leq \phi$.

Then, according to (2) and Remark 2 above, the objective functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$ with L2 regularizer are, respectively, defined as follows:

$$L_D(w) = \frac{1}{n}\sum_{j=1}^{n} L(w^T z_j) + \gamma\,||w||_2^2 \tag{13}$$

$$L_{D^{\setminus i}}(w) = \frac{1}{n}\sum_{j=1, j\neq i}^{n} L(w^T z_j) + \gamma\,||w||_2^2. \tag{14}$$

Theorem 1: Consider the FREL with L2 regularizer and a given training set D. Let D contain n input samples $x_i \in \mathbb{R}^d$ with its corresponding transformation z_i provided as in Remark 2, and that $||z_i||_2 \leq \phi$ $(i = 1, \ldots, n)$. Assume that the per-sample loss function $L(w^T z_i)$ in (5) is Lipchitz with constant δ. Let w_D and $w_{D^{\setminus i}}$ be the feature weighting results through minimizing the convex objective functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$ in (13) and (14), respectively. Then, FREL with L2 regularizer is uniformly weighting stable with stability bound $\beta = \delta\phi/n\gamma$.

Proof: Refer to Appendix A.

Remark 3: Theorem 1 shows that FREL with L2 regularizer has uniform weighting stability. Furthermore, the stability bound approaches zero as $O(1/n)$. Therefore, this is a tight bound.

Remark 4: To consider stability in the sense of Definition 2, for the case of removing q samples, the corresponding stability bound can be obtained similarly to that for a single sample removed and the bound is q times that of a single sample removed as well.

B. Stability Analysis for FREL With L1 Regularizer

Now, we turn to stability analysis for FREL with L1 regularizer. Due to the nature of the L1-norm, the feature selection algorithm with L1 regularizer usually results in sparse solutions, i.e., the feature vector output contains some elements that are zero. Xu *et al*. [26] proved that sparsity and stability are at odds with each other for classification and regression problems. They show that sparse algorithms are not stable, as defined in [23]. Specifically, if an algorithm encourages sparsity, then it is susceptible to small variations in input. They also proved that a sparse algorithm can identify redundant features (IRFs). Being IRF means that if the two features are highly dependent on each other, then removing one of the features would not affect the class-discriminative power of the algorithm. Therefore, a sparse algorithm may have nonunique optimal solutions and thus may be ill-posed. In this paper, we provide some constraints so that the stability of FREL with L1 regularizer in the sense of Definition 2 can be preserved.

For a given training set D from distribution $\mathbb{P}$, assume that there exists a true unique unknown feature weighting vector $w^\star$. Let w_D and $w_{D^{\setminus i}}$ be the optimal estimates of $w^\star$, respectively, where the optimality refers to that w_D and $w_{D^{\setminus i}}$ are optimal solutions as a result of minimizing the objective loss functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$, respectively

$$L_D(w) = \frac{1}{n}\sum_{j=1}^{n} L(w, x_j) + \gamma\,||w||_1 \tag{15}$$

$$L_{D^{\setminus i}}(w) = \frac{1}{n}\sum_{j=1, j\neq i}^{n} L(w, x_j) + \gamma\,||w||_1. \tag{16}$$

Then, we have

$$\begin{aligned} ||w_D - w_{D^{\setminus i}}||_2 &= ||w_D - w^\star - (w_{D^{\setminus i}} - w^\star)||_2 \\ &\leq ||w_D - w^\star||_2 + ||w_{D^{\setminus i}} - w^\star||_2. \end{aligned} \tag{17}$$

Let $||\triangle w_1||_2 = ||w_D - w^\star||_2$ and $||\triangle w_2||_2 = ||w_{D^{\setminus i}} - w^\star||_2$. Then

$$||w_D - w_{D^{\setminus i}}||_2 \leq ||\triangle w_1||_2 + ||\triangle w_2||_2. \tag{18}$$

To carry on the discussion of uniform weighting stability for FREL with L1 regularizer, we define the exactly sparse model below.

Definition 3: If some feature weights in the feature weighting vector are exactly zero, then this feature weighting model is *exactly sparse.*

Moreover, to analyze the stability of FREL with L1 regularizer, we also need some additional conditions, such as the sample-averaged loss functions $\mathcal{J}_D(w) = 1/n\sum_{j=1}^{n} L(w, x_j)$ and $\mathcal{J}_{D^{\setminus i}}(w) = 1/n\sum_{j=1, j\neq i}^{n} L(w, x_j)$ as in (3) are differentiable and they satisfy the strong convexity condition [27] defined below.

Definition 4: The sample-averaged loss function $\mathcal{J}_D(w)$ has the *strong convexity* with parameter $\kappa_D \geq 0$, if

$$\begin{aligned} \mathcal{J}_D(w^\star + \triangle w_1) - \mathcal{J}_D(w^\star) \geq\ & \langle \triangledown \mathcal{J}_D(w^\star), \triangle w_1 \rangle \\ & + \kappa_D ||\triangle w_1||_2^2 \end{aligned} \tag{19}$$

where $\langle , \rangle$ is the inner product. Similarly, we can define strong convexity for $\mathcal{J}_{D^{\setminus i}}(w)$ on $\triangle w_2$ with parameter $\kappa_{D^{\setminus i}} \geq 0$.

Remark 5: Refer to Remark 1 where we elaborated on some per-sample loss functions for $L(w, x_j)$. Among these functions, the log, square-square, and square-exponential losses are differentiable. Therefore, the corresponding sample-averaged loss functions $\mathcal{J}_D(w)$ and $\mathcal{J}_{D^{\setminus i}}(w)$ are also differentiable.

Remark 6: For those differentiable per-sample loss functions in Remark 5, it is evident that the square-exponential and square-square losses are strongly convex. Just as introduced

in [28], the log loss is also strongly convex. Then, their corresponding sample-averaged loss functions $\mathcal{J}_D(w)$ and $\mathcal{J}_{D^{\setminus i}}(w)$ are also strongly convex.

Theorem 2: Consider FREL with L1 regularizer and a given training set D from distribution $\mathbb{P}$. Let D contain n input samples $x_i \in \mathbb{R}^d$ $(i = 1, \ldots, n)$ and $w^\star$ be the true unique unknown feature weighting vector, and it is exactly sparse. Assume the sample-averaged loss function $\mathcal{J}_D(w)$ and $\mathcal{J}_{D^{\setminus i}}(w)$ are differentiable and have the strong convexity with parameter $\kappa_D \geq 0$ and $\kappa_{D^{\setminus i}} \geq 0$ as in Definition 4, respectively. Let w_D and $w_{D^{\setminus i}}$ be the sparse feature weighting results from minimizing the convex objective functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$ in (15) and (16), respectively. Then, for parameter $\gamma \geq \max[||\bigtriangledown \mathcal{J}_D(w^\star)||_\infty, ||\bigtriangledown \mathcal{J}_{D^{\setminus i}}(w^\star)||_\infty]$, FREL with L1 regularizer is uniformly weighting stable with stability bound $\beta = 2\sqrt{d}\gamma\,(1/\kappa_D + 1/\kappa_{D^{\setminus i}})$.

Proof: Refer to Appendix B.

Remark 7: For FREL with L1 regularizer, if its output is exactly sparse and the sample-averaged loss functions are strongly convex, then the feature weighting stability bound is inversely affected by the strong convexity constants κ_D and $\kappa_{D^{\setminus i}}$.

Remark 8: The bound also scales with the regularization parameter γ. This makes sense since the more sparse solutions lead to less feature weighting stability properties.

Remark 9: Consider stability in Definition 2 for the case of removing q samples. Let the corresponding sample-averaged loss function be strongly convex with parameter $\kappa_{D^{\setminus\setminus q}} \geq 0$. Then, the stability bound for q removed samples can be obtained similarly to that for a single sample remove, and the stability bound is $2\sqrt{d}\gamma\,(1/\kappa_D + 1/\kappa_{D^{\setminus\setminus q}})$.

C. Stability for Ensemble FREL

Based on Definition 2, the uniform weighting stability of the ensemble FREL proposed in Section II-D is defined as follows.

Definition 5: For a given training data set $D = \{x_i, y_i\}_{i=1}^n$ and any $i \in \{1, \ldots, n\}$, the ensemble FREL is *uniformly weighting stable* with stability bound $\beta_{\mathcal{E}}$, if

$$\left\| \mathbb{E}_{r_1,\ldots,r_m}\left[\frac{1}{m}\sum_{t=1}^{m} w_{D(r_t)}\right] - \mathbb{E}_{r_1,\ldots,r_m}\left[\frac{1}{m}\sum_{t=1}^{m} w_{D^{\setminus i}(r_t)}\right]\right\|_2 \leq \beta_{\mathcal{E}} \quad (20)$$

where $w_{D(r_t)}$ is the base feature weighting result of FREL on bootstrap subset $D(r_t)$ whose size is $\lceil \alpha n \rceil$ $(0 < \alpha < 1)$, and $w_{D^{\setminus i}(r_t)}$ is the base feature weighting result of FREL on bootstrap subset $D(r_t)$ with x_i, $i \in \{1, \ldots, n\}$, removed, m is the number of bootstrap subsets, $\mathbb{E}$ is the expectation, and for $t = 1, \ldots, m$, $r_t = \{r_t(1), r_t(2), \ldots, r_t(\lceil \alpha n \rceil)\}$ is an index sequence randomly drawn from the natural sequence $\{1, \ldots, n\}$ without replacement.

Theorem 3: Consider ensemble FREL described in Algorithm 2 and a given data set D containing n input samples x_i $(i = 1, \ldots, n)$. Bootstrap strategy is adopted with the sampling parameter α $(0 < \alpha < 1)$ to create m bootstrap subsets $D(r_t)$ with size $\lceil \alpha n \rceil$ for $t = 1, \ldots, m$, where $r_t = \{r_t(1), r_t(2), \ldots, r_t(\lceil \alpha n \rceil)\}$ is an index sequence randomly drawn from the natural sequence $\{1, \ldots, n\}$ without replacement, and $r_1, \ldots, r_m$ are i.i.d. Let β be the uniform stability bound of the base feature weighting algorithm FREL. Then, ensemble FREL is uniformly weighting stable with stability bound $\beta_{\mathcal{E}} \leq \alpha\beta$.

Proof: Refer to Appendix C.

Remark 10: Theorem 3 indicates that ensemble feature weighting has tighter stability bound than its base feature weighting, which is consistent with observations from experiments in [4] and [6] that ensemble strategy usually improves feature selection stability.

Remark 11: For the case of removing q samples, the corresponding stability bound can be obtained similarly to that for a single removed sample and the bound is approximately $\alpha^q\beta_{\setminus\setminus q}$, where $\beta_{\setminus\setminus q}$ is the stability bound for the base feature weighting with q samples removed.

IV. Experiments

In this section, we evaluate stability and accuracy of (ensemble) FREL in comparisons with some popular feature weighting algorithms. Four real life problems are considered.

The HDSSS problem is among the most challenging problems for feature selection, particularly if output stability is desired. To evaluate and illustrate algorithm accuracy and stability of FREL for HDSSS problems, we analyzed real-world microarray data including TOX, SMK, leukemia [29], and prostate [30]. The first two data sets are downloaded from http://featureselection.asu.edu/datasets.php, while the later two are available in [29] and [30]. The goal of analyzing these four HDSSS problems is to identify those genetic expressions that are linked to respective diseases.

The TOX data set contains 171 instances with 5748 genes. They consist of myocarditis and dilated cardiomyopathy (DCM) infected males and females as well as uninfected males and females. The DCM is often caused by viral infections and can occur more frequently in men than women. DCM infection increases a person's risk of dying from heart failure.

The SMK data set contains 187 smokers either with or without lung cancer. The total number of genes to be tested is 19 993.

Leukemias are primary disorders of the bone marrow. They are malignant neoplasms of hematopoietic stem cells. The leukemia data set to be analyzed includes 72 samples to be tested, which are from acute leukemia patients, either acute lymphoblastic leukemia or acute myelogenous leukemia. The total number of genes to be tested is 7129.

The prostate data set contains 136 samples of prostate cancer patients, which have 12 600 genes to be studied. Among the samples, 77 are tumor and 59 are normal.

As described, the four data sets (TOX, SMK, leukemia, and prostate) share the common traits of small samples (171, 187, 72, and 136) with an extremely high dimensionality in latent variables (5748, 19 993, 7129, and 12 600). They are typical HDSSS problems.

A. Algorithms for Comparison

For HDSSS problems, it is generally accepted that conventional feature selection algorithm should not be used directly for obtaining stable feature outputs [6], [7], [10]. Instead, ensemble algorithms are expected to improve stability properties of feature selection. We therefore focus on the ensemble FREL provided in Algorithm 2 when conducting comparisons in this paper. However, we still provide classification accuracy results using the original FREL presented in Algorithm 1.

For comparison purposes, we consider three specific FREL-based algorithms, as described in Algorithm 1: Log+L2, Log+L1, and Square+L2. The Log+L2 consists of log loss in (6) as per-sample loss function and L2 regularizer $||w||_2^2$ is used. For the Log+L1 algorithm, the per-sample loss function is the same as in Log+L2, but L1 regularizer $||w||_1$ is used. The Square+L2 is based on square-square loss function in (7) and L2 regularizer. For a sample x_i, the margin θ in Square+L2 is set as the Manhattan distance between $N_{\mathcal{M}}(x_i)$ and $N_{\mathcal{H}}(x_i)$. The ensemble versions of these three algorithms according to Algorithm 2 are named as En-Log+L2, En-Log+L1, and En-Square+L2, respectively.

Since the focus of this paper is on feature weighting, we therefore chose some popular feature weighting algorithms for comparison. The algorithms include Relief, ReliefF [31]–[33], Fisher score [34], En-Relief [4], En-ReliefF, En-Fisher, and variance reduction (VR)-Lmba. En-Relief, En-ReliefF, and En-Fisher are the ensemble versions of Relief, ReliefF, and Fisher score, respectively. The implementations of the En-Relief, En-ReliefF, and En-Fisher are similar to the ensemble FREL as described in Algorithm 2 with Relief, ReliefF, and Fisher score in place of the FREL, respectively.

The Relief algorithm is considered one of the most successful feature selection algorithms due to its simplicity and effectiveness [35]. The key idea of Relief is to iteratively estimate feature weights according to their ability to discriminate between the neighboring samples. In each iteration, a sample x_i is randomly selected and then two NNs of x_i are found, one from the same class and the other from a different class. The weight of the pth feature is then updated based on the distance between x_i and its two NNs on the pth feature. The ReliefF is an extension of Relief by considering several NNs to deal with multiple class problems.

Here, we would like to highlight the relationship between FREL and Relief (ReliefF). First of all, both algorithms can be viewed as hypothesis-margin-based approaches [36]. However, the differences between the two algorithms are evident: 1) our proposed FREL is a systematic framework for stable feature selection based on energy-based learning and regularization. Many specific algorithms can be viewed as realizations of FREL; 2) regularizations are considered in FREL but not in Relief; 3) Relief directly calculates the feature weights based on margins (distances) while FREL obtains feature weights based on margin losses, and the loss functions can be selected from a pool of candidate functions; and 4) a stability analysis for FREL is provided for the first time.

Fisher score is one of the most widely used feature selection methods. The key idea of Fisher score is to find feature weights such that in the data space spanned by the weighted features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible. Then, the criterion for Fisher score prefers features that have similar values for the samples from the same class and different values for the samples from different classes. Feature weights are obtained by computing the deviation of each feature from its mean value on all classes.

The VR-Lmba is a nonensemble stable feature weighting algorithm. VR-Lmba uses a sample weighting strategy [9], [10] to improve the stability of feature weighting algorithm—Lmba [37]. The sample weighting strategy introduced in [9] and [10] is an effective approach to improve the stability for any feature selection methods. It assigns different weights to samples before performing feature selection with the aim of VR. In this paper, we combine the sample weighting strategy with Lmba to obtain a stable feature weighting algorithm VR-Lmba, as presented in [7]. The feature weighting algorithm—Lmba is derived from energy-based model without regularization. In other words, the objective function for Lmba is similar to that in (2) without $\mathcal{R}(w)$. The square-square loss mentioned in Remark 1 is used as per-sample loss function $L(w, x_i)$ in (5). Several NNs in the predefined range are considered in Lmba.

B. Stability Measurement

Since in almost all feature selection applications, the ultimate outputs of a feature selector should be a subset of features that are considered most prominent. Therefore, in the literature, one usually makes use of the corresponding feature ranks in place of the feature weights for performance evaluation. Under the same consideration to evaluate the stability of FREL in this paper, we first compute the feature weights as outputs of FREL, then these weights are turned into feature ranks, as in [4] and [7]. Therefore, our stability measure used in this paper is based on feature ranks.

To be statistically significant, several different sets of feature ranking results are obtained to empirically compute the stability measure. We therefore use the bootstrap-based strategy without replacement. Consider the data set D with n instances and d features. Then, c sampling subsets $D(r_l)$, $l = 1, \ldots, c$, of size $\lceil \mu n \rceil$ $(0 < \mu < 1)$ are drawn randomly from D based on bootstrap sampling without replacement. Note that, we name the sampling subset $D(r_l)$ as a sample subset to distinguish it from a bootstrap subset in ensemble feature weighting procedure. Subsequently, feature weighting algorithms are performed on each of the c sample subsets. Just as emphasized earlier, the feature weighting results should be transformed to feature ranking results to calculate the stability measure. As such, each algorithm will result in c feature ranking results $\{R_1, R_2, \ldots, R_c\}$ on c sample subsets. For nonensemble feature weighting algorithm, such as VR-Lmba, Log+L2, and so on, to transform its feature weighting result on each sample subset into feature ranking result, the rank value for a feature is determined as follows. The best feature with the largest weight is assigned rank 1, and the worst rank d.

For ensemble feature weighting, feature ranking is obtained as described below.

Similar to the ensemble procedure described in Algorithm 2, each sample subset $D(r_l)$ with size $s = \lceil \mu n \rceil$ $(0 < \mu < 1)$, is still randomly sampled using bootstrap strategy without replacement to produce m bootstrap subsets $D(r_{lt})$ $(t = 1, \ldots, m)$ with size $\lceil \alpha s \rceil$ $(0 < \alpha < 1)$, and feature weighting algorithms (Relief, ReliefF, Fisher score, and our proposed FREL), are performed on each bootstrap subset $D(r_{lt})$ $(t = 1, \ldots, m)$ to obtain m base feature weighting vectors $\{w_{D(r_{l1})}, \ldots, w_{D(r_{lm})}\}$. To obtain the ensemble feature ranking result for the ensemble feature weighting algorithm on each sample subset $D(r_l)$, m base feature weighting vectors are correspondingly transformed to m base feature ranking vectors $\{v_{D(r_{l1})}, \ldots, v_{D(r_{lm})}\}$ based on the assignment rule above. The final feature rank for each sample subset $D(r_l)$, $l = 1, \ldots, c$, is obtained by averaging over the respective bootstrapping subset-based feature ranks, i.e., $R_l = 1/m \sum_{t=1}^{m} v_{D(r_{lt})}$ [4].

Consider a feature ranking vector set $\{R_1, R_2, \ldots, R_c\}$, where $R_l = (R_l^1, R_l^2, \ldots, R_l^d)$, $l = 1, 2, \ldots, c$, is the feature ranking result for the d features on the lth sample subset. Feature selection stability is measured by comparing similarities among the feature outputs on the c sample subsets. The more similar the outputs are, the higher the stability measure is. The overall stability is defined as the average similarity based on all pairwise similarities between different feature ranking results

$$R_{\text{sta}} = \frac{2}{c(c-1)} \sum_{l=1}^{c-1} \sum_{l'=l+1}^{c} \text{Sim}(R_l, R_{l'}) \tag{21}$$

where $\text{Sim}(R_l, R_{l'})$ represents a similarity measure between feature ranking results R_l and $R_{l'}$. For feature ranking, the Spearman rank correlation coefficient [4], [38] is used to calculate the similarity

$$\text{Sim}(R_l, R_{l'}) = 1 - 6 \sum_{p=1}^{d} \frac{(R_l^p - R_{l'}^p)^2}{d(d^2-1)}. \tag{22}$$

C. Experiments Performed for Stability

Based on the stability measurement procedure described in Section IV-B, for a given data set, $c = 10$ sample subsets containing $\mu = 90\%$ of the data are randomly drawn without replacement using bootstrap-based strategy. This percentage was chosen as in [4] to assess stability with respect to relatively small changes in the data set. For example, leukemia data set is randomly drawn using bootstrap-based strategy without replacement to create 10 sample subsets, thus each sample subset contains 64 patient samples with 7129 genes.

Then, for each sample subset, ensemble feature weighting algorithms (En-Log+L2 with $\gamma = 1$, En-Log+L1 with $\gamma = 0.01$, En-Square+L2 with $\gamma = 0.1$, En-Relief, En-ReliefF with 10 NNs, and En-Fisher) are applied as described in Section IV-B with $\alpha = 0.9$ to obtain feature ranking results. Simultaneously, the nonensemble feature weighting algorithm VR-Lmba, which is another method for improving the stability of feature selection, is also applied to each sample subset. For 10 sample subsets, we obtain 10 feature ranking results for each feature weighting algorithm. For each feature weighting algorithm, the similarity between feature ranking result pairs is calculated using Spearman rank correlation coefficient in (22). The stability of each feature weighting algorithm is thus the average similarity over all pairwise similarities calculated by (21).

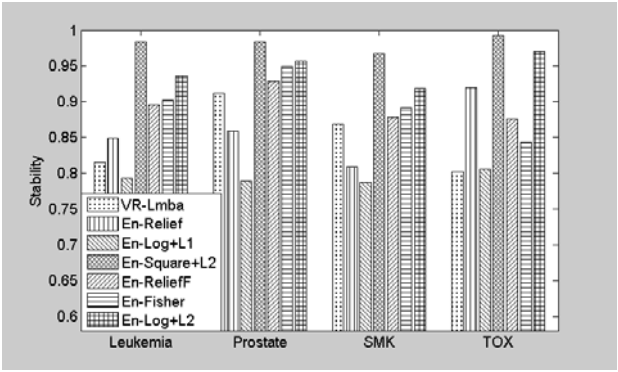


Fig. 1. Experimental evaluation of stability for $m = 20$, the number of base feature selectors, for each of the seven candidate stable feature selection algorithms.

Moreover, we examine the effect of the regularization parameter γ in (2) on the stability of our FREL. As examples, one original algorithm derived from FREL, i.e., Log+L2, and one ensemble algorithm derived from ensemble FREL, i.e., En-Log+L2 with $m = 20$, are chosen.

D. Experimental Results for Stability

We first examine the effect of the number of bootstrap subsets used in ensemble methods, namely, how m affects the stability measure. We see that all algorithms display an upward trend in stability as m increases, but saturates at around $m = 20$. Since VR-Lmba is not an ensemble method, its stability remains constants. The stability results for ensemble feature weighting algorithms with $m = 20$ and VR-Lmba are therefore shown in Fig. 1. From Fig. 1, we observe that the proposed ensemble FREL with L2 regularizer (En-log+L2 and En-Square+L2), always have the highest stability among all the algorithms. Our algorithm with L1 regularizer (En-log+L1), has the lowest stability, which is consistent with the observation in [26] that sparsity and stability are at odds with each other.

On the other hand, we examine the effect of the regularization parameter γ in (2) on the stability of FREL (Log+L2 and En-Log+L2). Results on leukemia and prostate data sets are included. Similar results were obtained for the other two data sets. They are not included here due to space limitations. The experimental results are shown in Fig. 2. We observe that along with the increase in γ, the stability of Log+L2 and En-Log+L2 are improved, which is consistent with our theoretical analysis.

E. Experiments Performed for Accuracy

A good feature selector has to be both stable and accurate. Once stable features are selected, an important consideration

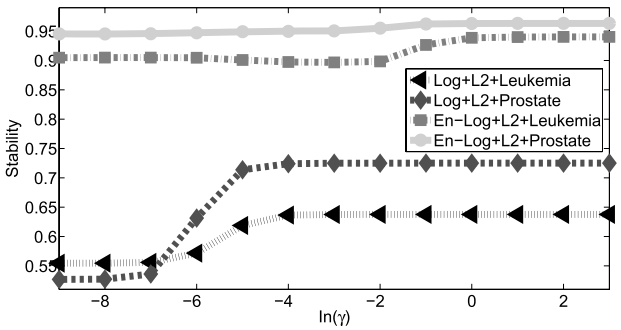


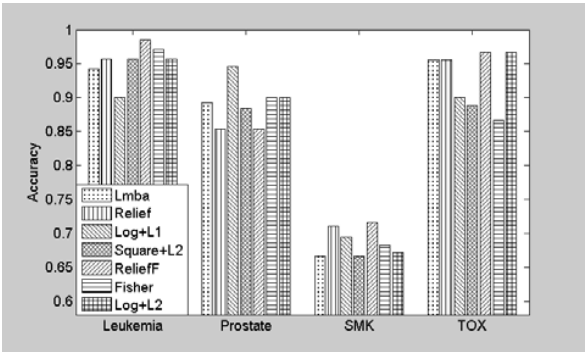
Fig. 2. Experimental evaluation of stability as a function of γ, the regularization parameter in (2), for Log+L2 and En-Log+L2 on leukemia and prostate.

in many applications is the classification accuracy using the selected features. The accuracy has to be evaluated in conjunction with a classification model based on the selected features. In experiments conducted in this paper, 1-NN (1NN) classifier, 3-NNs (3NN) classifier, the linear support vector machine (SVM) with $C = 1$ [39], and SVM with polynomial kernel are used as classification models since they are generally considered easy to apply and good classifiers. Each classifier is used for classification based on each of the feature weighting algorithms introduced in Section IV-A. Classification accuracy is assessed using a 10-fold cross-validation. For each fold, a feature weighting algorithm was applied to the training data to obtain the feature ranking result. For ensemble feature weighting algorithms (En-log+L2, En-log+L1, En-Square+L2, En-Relief, En-ReliefF, and En-Fisher), based on the experimental results in Section IV-D, for each fold, $m = 20$ bootstrap subsets of training data were randomly drawn with $\alpha = 0.9$ to create the ensemble feature weighting algorithms. After the features are ranked in descending order, different numbers of important features are selected with top ranks one by one to create classifiers. Note that it is often observed in microarray data that only a small amount ($\approx$50) of features, i.e., genes, is relevant [9], [10], [16], thus the number of selected important features from ranking results is less than 100 in our experiments. Once test result based on testing data is obtained for each fold, the final classification accuracy results are obtained by averaging over the 10 folds.

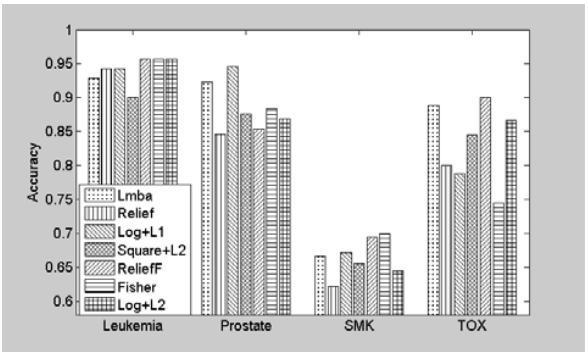
F. Experimental Results for Accuracy

Our accuracy results are provided for the case of using 50 selected features. Fig. 3 shows a summary of the accuracy values for regular feature selection algorithms (Lmba, Relief, ReliefF with 10 NNs, Fisher score, Log+L2, Log+L1, and Square+L2) using 1NN, 3NN, linear SVM, and polynomial kernel SVM classifier. Fig. 4 shows a summary of accuracy values for algorithms designed to improve stability (En-Relief, En-ReliefF, En-Fisher, En-Log+L2, En-Log+L1, En-Square+L2, and VR-Lmba) using 1NN, 3NN, linear SVM, and polynomial kernel SVM classifier.

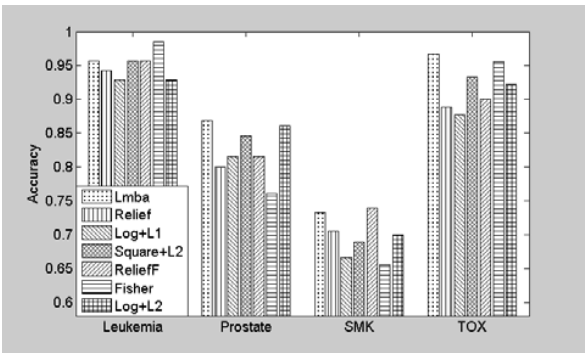
These results as summarized in Figs. 3 and 4 show that no one algorithm is consistently better than any other on the four tested data sets. However, FREL and ensemble FREL are comparable with other algorithms most of the time.



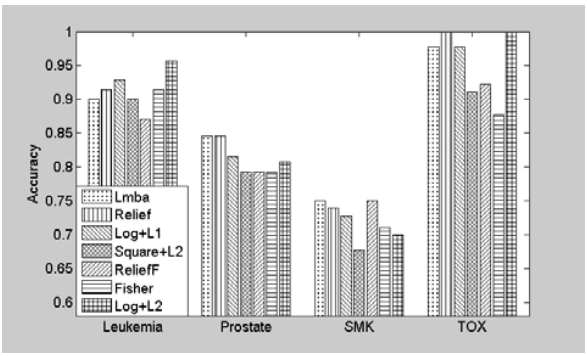
(a)



(b)



(c)

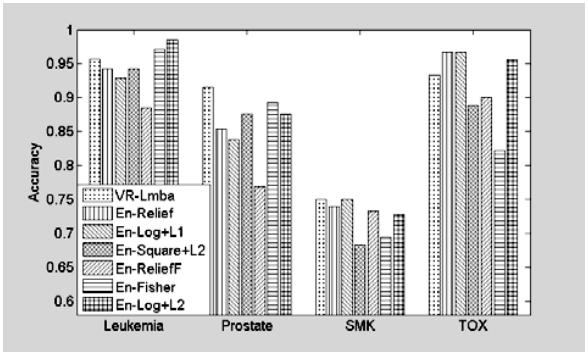


(d)

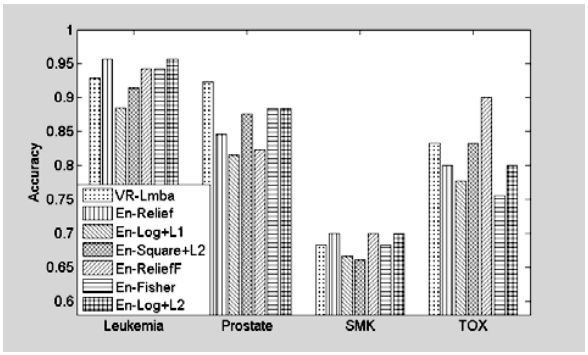
Fig. 3. Experimental results for accuracy of original feature selection methods using different classifiers. (a) 1NN. (b) 3NN. (c) Linear SVM. (d) Polynomial SVM.

G. Evaluation of Tradeoff Between Stability and Accuracy

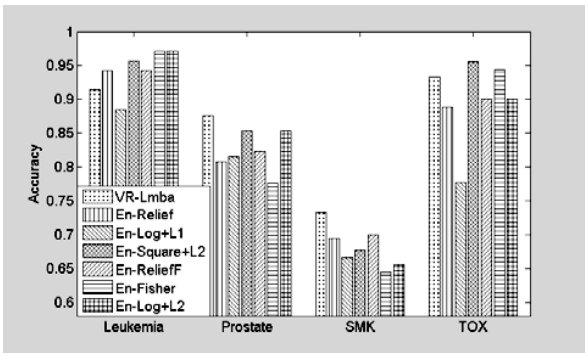
To measure the tradeoff between stability and classification accuracy of a feature selection algorithm, we take reference of the robustness-performance tradeoff in [4] to



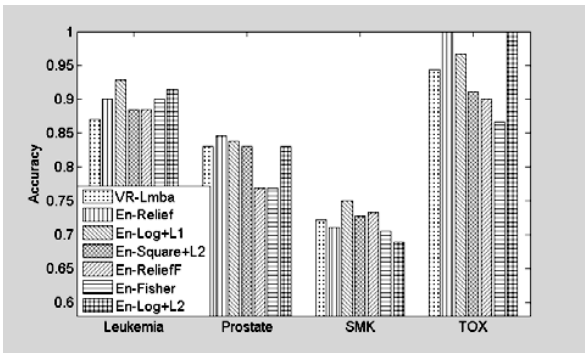
(a)



(b)

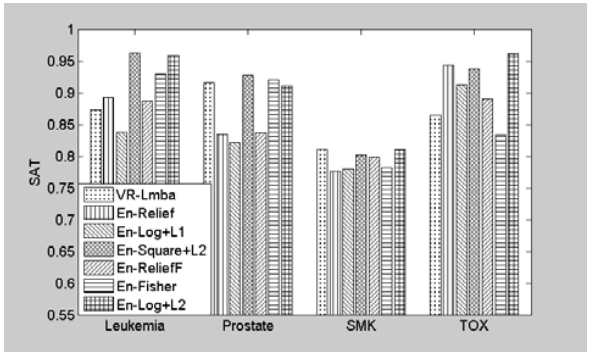


(c)

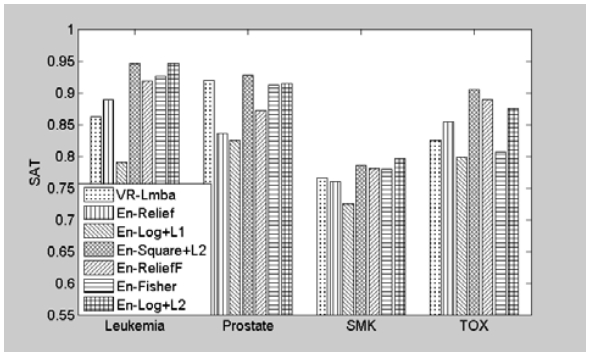


(d)

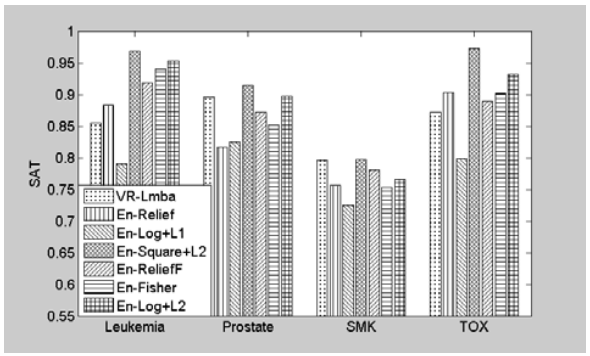
Fig. 4. Experimental results for accuracy of methods designed for stable feature selection using different classifiers. (a) 1NN. (b) 3NN. (c) Linear SVM. (d) Polynomial SVM.



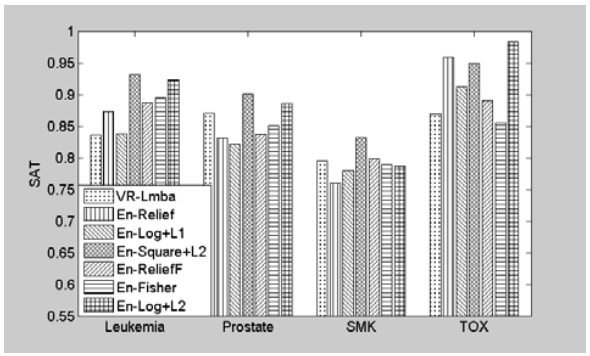
(a)



(b)



(c)



(d)

Fig. 5. Experimental results about tradeoff between stability and accuracy for methods designed for stable feature selection using different classifiers. (a) 1NN. (b) 3NN. (c) Linear SVM. (d) Polynomial SVM.

measure the tradeoff between feature stability and classification accuracy in this paper. Specifically, we define stability-accuracy trade-off (SAT) as SAT $=$ (2 $\times$ stability $\times$ accuracy)/(stability $+$ accuracy) where stability can be implemented by R_{sta} in (21), accuracy is evaluated using the classification outcome based on the selected features. The stability value for ensemble feature weighting when $m = 20$ and the corresponding accuracy when the number of selected feature is

50 are used to calculate their tradeoff. The experimental results for different classifiers are shown in Fig. 5 where ensemble FREL with L2 regularizer is shown providing a better tradeoff between stability and accuracy than other compared methods.

V. Conclusion

In this paper, we have proposed a new framework for FREL, which includes many useful stable feature weighting algorithms as its realizations. We also provide for the first time the theoretical results for the uniform weighting stability of FREL with L1 and L2 regularizers. Ensemble FREL is introduced as a means of further improvement of stability, the stability of which is also provided. To evaluate FREL and ensemble FREL performance, we make use of three specific realizations, Log+L1, Log+L2, and Square+L2, respectively. Several other popular feature selection algorithms are included in comparison with benchmark performances based on challenging HDSSS problems. Our experimental results show that our ensemble FREL when using the L2 regularizer outperforms other algorithms in stability while providing comparable classification accuracy.

Appendix A
Proof of the Theorem 1

Proof: Let $\triangle w_D = w_D - w_{D^{\setminus i}}$, where w_D and $w_{D^{\setminus i}}$ are the feature weighting results through minimizing the convex objective functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$ in (13) and (14), respectively. As such $L_D(w_D)$ and $L_{D^{\setminus i}}(w_{D^{\setminus i}})$ retain minimum values at w_D and $w_{D^{\setminus i}}$, respectively. Accordingly, this leads to the following for any $a \in [0, 1]$:

$$L_D(w_D) - L_D(w_D - a \triangle w_D) \leq 0 \tag{23}$$

$$L_{D^{\setminus i}}(w_{D^{\setminus i}}) - L_{D^{\setminus i}}(w_{D^{\setminus i}} + a \triangle w_D) \leq 0. \tag{24}$$

Equations (13) and (14) are used to replace the corresponding terms in (23) and (24). It is evident that $1/n \sum_{j=1}^{n} L(w_D^T z_j) = 1/n \sum_{j=1, j\neq i}^{n} L(w_D^T z_j) + 1/n L(w_D^T z_i)$. We use a shorthand notation $1/n \sum_{j^{\setminus i}}$ for $1/n \sum_{j=1, j\neq i}^{n}$ for the ease of discussion. Then, using (23) and (24) together, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j^{\setminus i}} L(w_D^T z_j) + \frac{1}{n} L(w_D^T z_i) \\
&\quad - \frac{1}{n}\sum_{j^{\setminus i}} L((w_D - a \triangle w_D)^T z_j) \\
&\quad - \frac{1}{n} L((w_D - a \triangle w_D)^T z_i) + \frac{1}{n}\sum_{j^{\setminus i}} L(w_{D^{\setminus i}}^T z_j) \\
&\quad - \frac{1}{n}\sum_{j^{\setminus i}} L((w_{D^{\setminus i}} + a \triangle w_D)^T z_j) + \gamma . ||w_D||_2^2 \\
&\quad - \gamma \, ||w_D - a \triangle w_D||_2^2 + \gamma \, ||w_{D^{\setminus i}}||_2^2 \\
&\quad - \gamma \, ||w_{D^{\setminus i}} + a \triangle w_D||_2^2 \\
&\leq 0.
\end{aligned} \tag{25}$$

Since the per-sample loss function in (5) is convex, then by Jensen's inequality

$$\begin{aligned}
L((w_D - a \triangle w_D)^T z_j) &= L((1-a) w_D^T z_j + a w_{D^{\setminus i}}^T z_j) \\
&\leq (1-a) L(w_D^T z_j) + a L(w_{D^{\setminus i}}^T z_j) \\
&= L(w_D^T z_j) \\
&\quad - a(L(w_D^T z_j) - L(w_{D^{\setminus i}}^T z_j)).
\end{aligned} \tag{26}$$

Similarly, we also can obtain

$$\begin{aligned}
L((w_{D^{\setminus i}} + a \triangle w_D)^T z_j)) \leq L(w_{D^{\setminus i}}^T z_j) + a(L(w_D^T z_j) \\
- L(w_{D^{\setminus i}}^T z_j)).
\end{aligned} \tag{27}$$

Substituting two identities (26) and (27) into (25) leads to

$$\begin{aligned}
&||w_D||_2^2 - ||w_D - a \triangle w_D||_2^2 - ||w_{D^{\setminus i}} + a \triangle w_D||_2^2 \\
&\quad + ||w_{D^{\setminus i}}||_2^2 \\
&\leq \frac{a}{n\gamma}(L(w_{D^{\setminus i}}^T z_i) - L(w_D^T z_i)).
\end{aligned} \tag{28}$$

Note that the inequality $L(w_{D^{\setminus i}}^T z_i) - L(w_D^T z_i) \leq \delta |w_{D^{\setminus i}}^T z_i - w_D^T z_i|$ holds because the per-sample loss function $L(w^T z_i)$ is Lipchitz with δ [40]. Therefore

$$\begin{aligned}
&||w_D||_2^2 - ||w_D - a \triangle w_D||_2^2 - ||w_{D^{\setminus i}} + a \triangle w_D||_2^2 \\
&\quad + ||w_{D^{\setminus i}}||_2^2 \leq \frac{a\delta}{n\gamma} |w_{D^{\setminus i}}^T z_i - w_D^T z_i| \\
&= \frac{a\delta}{n\gamma} |\triangle w_D^T z_i|.
\end{aligned} \tag{29}$$

If we set $a = 1/2$, the left-hand side of (29) amounts to

$$\begin{aligned}
&||w_D||_2^2 + ||w_{D^{\setminus i}}||_2^2 - \frac{1}{2}||w_D + w_{D^{\setminus i}}||_2^2 \\
&\quad = \frac{1}{2}||w_D||_2^2 + \frac{1}{2}||w_{D^{\setminus i}}||_2^2 - w_D^T w_{D^{\setminus i}} \\
&\quad = \frac{1}{2}||w_D - w_{D^{\setminus i}}||_2^2 \\
&\quad = \frac{1}{2}||\triangle w_D||_2^2.
\end{aligned} \tag{30}$$

Thus

$$||\triangle w_D||_2^2 \leq \frac{\delta}{n\gamma} |\triangle w_D^T z_i|. \tag{31}$$

Then, based on the Cauchy–Schwarz inequality, we have

$$|\triangle w_D^T z_i| \leq ||\triangle w_D||_2 ||z_i||_2. \tag{32}$$

Combining (31) and (32) above, and using $||z_i||_2 \leq \phi$, we obtain the uniform stability bound for FREL with L2 regularizer

$$||w_D - w_{D^{\setminus i}}||_2 = ||\triangle w_D||_2 \leq \frac{\delta\phi}{n\gamma}. \tag{33}$$

■

APPENDIX B
PROOF OF THE THEOREM 2

Proof: Let w_D and $w_{D^{\setminus i}}$ be the optimal estimates of $w^\star$, respectively, where the optimality refers to that w_D and $w_{D^{\setminus i}}$ are optimal solutions as a result of minimizing the objective loss functions $L_D(w)$ and $L_{D^{\setminus i}}(w)$ in (15) and (16), respectively. We have $||w_D - w_{D^{\setminus i}}||_2 \leq ||\triangle w_1||_2 + ||\triangle w_2||_2$, as in (18).

To analyze the two terms $||\triangle w_1||_2$ and $||\triangle w_2||_2$ in (18), we start with the decomposability property of L1 regularizer below.

In consideration of exact sparsity as defined in Definition 3, the feature weighting results are d-dimensional vectors $\{w(1), w(2), \ldots, w(d)\}$ with some weights being exactly zero. Suppose the number of features with nonzero weights is b. Let S be an index set whose b components correspond to the index of those features with nonzero weights. For example, $S = \{1, 2\}$ indicates that those weights from $w(3)$ through $w(d)$ are zeros while $b = 2$. Let $\mathbb{R}^d$ be the d-dimension real space. Then, we define the subspace M as

$$M := \{\sigma \in \mathbb{R}^d \quad | \quad \sigma_p = 0 \quad \text{for all} \quad p \notin S\}. \tag{34}$$

The orthogonal complement subspace of M is

$$M^\perp := \{\psi \in \mathbb{R}^d \quad | \quad \psi_p = 0 \quad \text{for all} \quad p \in S\}. \tag{35}$$

Remark 12: Subspace M is the model subspace capturing the constraints specified by the L1 regularizer in $L_D(w)$ or $L_{D^{\setminus i}}(w)$, while $M^\perp$ is the orthogonal complement subspace of M, and it is considered a perturbation subspace deviating away from the model subspace M. Then, $M \bigcup M^\perp = \mathbb{R}^d$.

We are ready to define the decomposability property of L1 regularizer with respect to the model subspace and its orthogonal complement subspace as in [41]–[43].

Definition 6: Given a subspace pair M and $M^\perp$ as defined in (34) and (35), respectively, an L1 regularizer is *decomposable* with respect to $(M, M^\perp)$, such that

$$||\sigma + \psi||_1 = ||\sigma||_1 + ||\psi||_1 \tag{36}$$

for all $\sigma \in M$ and $\psi \in M^\perp$.

Next, we analyze the bound of $||\triangle w_1||_2$ in (18). To simplify the notation, we drop the subscript of $\triangle w_1$ and use $\triangle w$ instead in this proof. Let

$$\mathcal{F}(\triangle w) = L_D(w^\star + \triangle w) - L_D(w^\star). \tag{37}$$

Since $w_D = w^\star + \triangle w$ is the minimizer of $L_D(w)$ in (15), then $\triangle w$ must satisfy $\mathcal{F}(\triangle w) \leq 0$.

If $L_D(w)$ is replaced by the right-hand side of (15), $\mathcal{F}(\triangle w)$ is then changed as

$$\begin{aligned}\mathcal{F}(\triangle w) = {} & \mathcal{J}_D(w^\star + \triangle w) - \mathcal{J}_D(w^\star) \\ & + \gamma\,(||w^\star + \triangle w||_1 - ||w^\star||_1).\end{aligned} \tag{38}$$

Note that the function $\mathcal{F}(\triangle w)$ consists of two differences: one is between the sample-averaged loss functions, i.e., $\mathcal{J}_D(w^\star + \triangle w) - \mathcal{J}_D(w^\star)$, and the other is between the regularizers, i.e., $||w^\star + \triangle w||_1 - ||w^\star||_1$.

Consider first the difference between regularizers $||w^\star + \triangle w||_1 - ||w^\star||_1$. Based on the subspaces M and $M^\perp$ defined above, it is evident that $w^\star = w^\star_M + w^\star_{M^\perp}$, where $w^\star_M$ and $w^\star_{M^\perp}$ are the projections of $w^\star$ onto subspace M and its orthogonal complement subspace $M^\perp$, respectively. The projection operation is defined as

$$w^\star_M = \Pi_M(w^\star) := \text{argmin}_{u \in M} ||w^\star - u||_2. \tag{39}$$

Similarly, we can obtain $\triangle w = \triangle w_M + \triangle w_{M^\perp}$, where $\triangle w_M$ and $\triangle w_{M^\perp}$ are the projections of $\triangle w$ onto subspaces M and $M^\perp$, respectively. The definitions for $w^\star_{M^\perp}$, $\triangle w_M$, and $\triangle w_{M^\perp}$ are given in an analogous manner to $w^\star_M$.

Then, by the triangle inequality, we have $||w^\star||_1 = ||w^\star_M + w^\star_{M^\perp}||_1 \leq ||w^\star_M||_1 + ||w^\star_{M^\perp}||_1$ and

$$\begin{aligned}||w^\star + \triangle w||_1 &= ||w^\star_M + w^\star_{M^\perp} + \triangle w_M + \triangle w_{M^\perp}||_1 \\ &\geq ||w^\star_M + \triangle w_{M^\perp}||_1 - ||w^\star_{M^\perp} + \triangle w_M||_1 \\ &\geq ||w^\star_M + \triangle w_{M^\perp}||_1 - ||w^\star_{M^\perp}||_1 \\ &\quad -||\triangle w_M||_1.\end{aligned} \tag{40}$$

By the decomposability property of L1 regularizer as in Definition 6, $||w^\star_M + \triangle w_{M^\perp}||_1 = ||w^\star_M||_1 + ||\triangle w_{M^\perp}||_1$ is obtained, so that $||w^\star + \triangle w||_1 \geq ||w^\star_M||_1 + ||\triangle w_{M^\perp}||_1 - ||w^\star_{M^\perp}||_1 - ||\triangle w_M||_1$. Therefore

$$\begin{aligned}||w^\star + \triangle w||_1 - ||w^\star||_1 \geq {} & ||\triangle w_{M^\perp}||_1 - ||\triangle w_M||_1 \\ & -2||w^\star_{M^\perp}||_1.\end{aligned} \tag{41}$$

For an exactly sparse model as in Definition 3, define a model subspace M containing $w^\star$, i.e., $w^\star \in M$, to guarantee $||w^\star_{M^\perp}||_1 = 0$ [41]–[43], and obtain

$$||w^\star + \triangle w||_1 - ||w^\star||_1 \geq ||\triangle w_{M^\perp}||_1 - ||\triangle w_M||_1. \tag{42}$$

Now, we turn to the difference between the sample-averaged loss functions $\mathcal{J}_D(w^\star + \triangle w) - \mathcal{J}_D(w^\star)$ in $\mathcal{F}(\triangle w)$, as defined in (38). To analyze the differences between the sample-averaged loss functions in $\mathcal{F}(\triangle w)$, we let the sample-averaged loss functions $\mathcal{J}_D(w)$ be differentiable and satisfy strong convexity as in Definition 4. Based on (42) and the strong convexity of $\mathcal{J}_D(w)$ in Definition 4, the function $\mathcal{F}(\triangle w)$ is expressed as

$$\begin{aligned}\mathcal{F}(\triangle w) \geq {} & \mathcal{J}_D(w^\star + \triangle w) - \mathcal{J}_D(w^\star) \\ & + \gamma\,(||\triangle w_{M^\perp}||_1 - ||\triangle w_M||_1) \\ \geq {} & \langle \nabla \mathcal{J}_D(w^\star), \triangle w \rangle + \kappa_D ||\triangle w||_2^2 \\ & + \gamma\,(||\triangle w_{M^\perp}||_1 - ||\triangle w_M||_1).\end{aligned} \tag{43}$$

By the Cauchy–Schwarz inequality application, we have

$$|\langle \nabla \mathcal{J}_D(w^\star), \triangle w \rangle| \leq ||\nabla \mathcal{J}_D(w^\star)||_\infty ||\triangle w||_1. \tag{44}$$

Without loss of generality, assume that

$$\gamma \geq ||\nabla \mathcal{J}_D(w^\star)||_\infty. \tag{45}$$

We can conclude that $\langle \nabla \mathcal{J}_D(w^\star), \triangle w \rangle \geq -\gamma ||\triangle w||_1$. Thus

$$\begin{aligned}\mathcal{F}(\triangle w) \geq {} & \kappa_D ||\triangle w||_2^2 + \gamma\,(||\triangle w_{M^\perp}||_1 - ||\triangle w_M||_1) \\ & -\gamma ||\triangle w||_1.\end{aligned} \tag{46}$$

By the triangle inequality, $||\triangle w||_1 = ||\triangle w_{M^\perp} + \triangle w_M||_1 \leq ||\triangle w_{M^\perp}||_1 + ||\triangle w_M||_1$, and hence

$$\mathcal{F}(\triangle w) \geq \kappa_D ||\triangle w||_2^2 - 2\gamma ||\triangle w_M||_1. \quad (47)$$

Note that, $||\triangle w_M||_1 \leq \sqrt{d}||\triangle w_M||_2$. Since the projection $\triangle w_M$ is defined similarly to (39) in terms of the L2-norm, it is nonexpansive. Since $0 \in M$

$$||\triangle w_M||_2 = ||\Pi_M(\triangle w) - \Pi_M(0)||_2 \leq ||\triangle w - 0||_2 = ||\triangle w||_2. \quad (48)$$

In the end, we obtain

$$\mathcal{F}(\triangle w) \geq \kappa_D ||\triangle w||_2^2 - 2\sqrt{d}\gamma ||\triangle w||_2. \quad (49)$$

As discussed above, $\mathcal{F}(\triangle w) \leq 0$

$$\kappa_D ||\triangle w||_2^2 - 2\sqrt{d}\gamma ||\triangle w||_2 \leq 0. \quad (50)$$

Then, we obtain

$$||\triangle w||_2 \leq \frac{2\sqrt{d}\gamma}{\kappa_D} \quad (51)$$

and we change notation $\triangle w$ back to $\triangle w_1$, then $||\triangle w_1||_2 \leq 2\sqrt{d}\gamma/\kappa_D$.

Now, we analyze the bound of $\triangle w_2$. Similarly, we let $\mathcal{F}(\triangle w_2)$ be

$$\mathcal{F}(\triangle w_2) = L_{D^{\setminus i}}(w^\star + \triangle w_2) - L_{D^{\setminus i}}(w^\star). \quad (52)$$

Similar to the analysis for $\mathcal{F}(\triangle w)$ above, we can obtain

$$||\triangle w_2||_2 \leq \frac{2\sqrt{d}\gamma}{\kappa_{D^{\setminus i}}} \quad (53)$$

with $\gamma \geq ||\nabla \mathcal{J}_{D^{\setminus i}}(w^\star)||_\infty$.

Based on (18) and without loss of generality, assume

$$\gamma \geq \max[||\nabla \mathcal{J}_D(w^\star)||_\infty, ||\nabla \mathcal{J}_{D^{\setminus i}}(w^\star)||_\infty] \quad (54)$$

we obtain the stability bound of feature weighting with L1 regularizer

$$||w_D - w_{D^{\setminus i}}||_2 \leq ||\triangle w_1||_2 + ||\triangle w_2||_2 \leq 2\sqrt{d}\gamma\left(\frac{1}{\kappa_D} + \frac{1}{\kappa_{D^{\setminus i}}}\right). \quad (55)$$

■

APPENDIX C
PROOF OF THE THEOREM 3

Proof: For the uniform stability of ensemble FREL in (20), the left side can be bounded by taking the L2 norm inside the expectation by Jensen's inequality. According to Jensen's inequality, let f be a convex function and x be a random variable. Then, $f(\mathbb{E}(x)) \leq \mathbb{E}(f(x))$. For our case, L2-norm is convex, so we obtain

$$\beta_{\mathcal{E}} = \left\| \mathbb{E}_{r_1,\ldots,r_m}\left[\frac{1}{m}\sum_{t=1}^{m} w_{D(r_t)} - \frac{1}{m}\sum_{t=1}^{m} w_{D^{\setminus i}(r_t)}\right]\right\|_2 \leq \mathbb{E}_{r_1,\ldots,r_m}\left[\left\|\frac{1}{m}\sum_{t=1}^{m} w_{D(r_t)} - \frac{1}{m}\sum_{t=1}^{m} w_{D^{\setminus i}(r_t)}\right\|_2\right]. \quad (56)$$

Since $r_1, \ldots, r_m$ are i.i.d. and suppose they have the same distribution as r, which models bootstrapping once, as in [24]. By the triangle inequality

$$\beta_{\mathcal{E}} \leq \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}_{r_t}[||w_{D(r_t)} - w_{D^{\setminus i}(r_t)}||_2] = \mathbb{E}_r[||w_{D(r)} - w_{D^{\setminus i}(r)}||_2] = \mathbb{E}_r[||\triangle w_{D(r)}||_2]. \quad (57)$$

Therefore, according to (57), the ensemble stability bound may now be considered similarly as in Definition 2 of removing a single sample x_i from the data set D. Since r is a sampled subset of $\{1, 2, \ldots, n\}$, we need to consider two possibilities of whether i belongs to r or not. To do so, we introduce an indicator function $\mathbb{I}(.)$. Note that if i is not in r, which means the sample x_i is not in the bootstrap subset $D(r)$, i.e., $D(r) = D^{\setminus i}(r)$, then the term $\mathbb{E}_r[||\triangle w_{D(r)}||_2 \mathbb{I}(i \notin r)] = 0$. We have the following:

$$\begin{aligned}\beta_{\mathcal{E}} &\leq \mathbb{E}_r[||\triangle w_{D(r)}||_2(\mathbb{I}(i \in r) + \mathbb{I}(i \notin r))] \\ &= \mathbb{E}_r[||\triangle w_{D(r)}||_2 \mathbb{I}(i \in r)] + \mathbb{E}_r[||\triangle w_{D(r)}||_2 \mathbb{I}(i \notin r)] \\ &= \mathbb{E}_r[||\triangle w_{D(r)}||_2 \mathbb{I}(i \in r)]. \end{aligned} \quad (58)$$

The size of bootstrap subset $D(r)$ is $\lceil \alpha n \rceil$ $(0 < \alpha < 1)$, then $\mathbb{E}_r(\mathbb{I}(i \in r)) = \lceil \alpha n \rceil / n \approx \alpha$ because this sampling is done without replacement, and $||\triangle w_{D(r)}||_2 \leq \beta$ due to the uniform stability of the base feature weight result, then we have

$$\beta_{\mathcal{E}} \leq \alpha\beta. \quad (59)$$

■

REFERENCES

[1] Z. Zhao, "Spectral feature selection for mining ultrahigh dimensional data," Ph.D. dissertation, Dept. School Comput., Informat., Decision Syst. Eng., Arizona State Univ., Phoenix, AZ, USA, 2010.

[2] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 494–502, Apr. 2005.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 31, pp. 1157–1182, Jan. 2003.

[4] Y. Saeys, T. Abeel, and Y. van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. 25th Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2008, pp. 313–325.

[5] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 567–575.

[6] T. Abeel, T. Helleputte, Y. van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[7] Y. Li, S. Gao, and S. Chen, "Ensemble feature weighting based on local learning and diversity," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1019–1025.

[8] A. Woznica, P. Nguyen, and A. Kalousis, "Model mining for robust feature selection," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 913–921.

[9] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," in *Proc. 10th Int. Conf. Data Mining*, Dec. 2010, pp. 206–215.

[10] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 1, pp. 262–272, Jan./Feb. 2012.

[11] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 803–811.

[12] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Comput. Biol. Chem.*, vol. 34, no. 4, pp. 215–225, 2010.
[13] Y. LeCun, S. Chopra, R. Hadsell, M. A. Ranzato, and F. J. Huang, *A Tutorial on Energy-Based Model*. Cambridge, MA, USA: MIT Press, 2006.
[14] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: Scalable online collaborative filtering," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 271–280.
[15] S. Dudoit, J. Fridlyand, and T. P. Spee, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
[16] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.
[17] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[18] B. Wang and H.-D. Chiang, "Elite: Ensemble of optimal input-pruned neural networks using TRUST-TECH," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 96–109, Jan. 2011.
[19] S. Tang, Y.-T. Zheng, Y. Wang, and T.-S. Chua, "Sparse ensemble learning for concept detection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 43–54, Feb. 2012.
[20] S. Gutta, J. R. J. Huang, P. Jonathon, and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," *IEEE Trans. Neural Netw.*, vol. 11, no. 4, pp. 948–960, Jul. 2000.
[21] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 531–542, Feb. 2010.
[22] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, Aug. 1996.
[23] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Jan. 2002.
[24] A. Elisseeff, T. Evgeniou, and M. Pontil, "Stability of randomized learning algorithm," *J. Mach. Learn. Res.*, vol. 6, pp. 55–79, Jan. 2005.
[25] S. Agarwal and P. Niyogi, "Generalization bounds for ranking algorithm via algorithmic stability," *J. Mach. Learn. Res.*, vol. 10, pp. 441–474, Feb. 2009.
[26] H. Xu, C. Caramanis, and S. Mannor, "Sparse algorithms are not stable: A no-free-lunch theorem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 187–193, Jan. 2012.
[27] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Boston, MA, USA: Kluwer, 2004.
[28] M. Tan, I. W. Tsang, and L. Wang, "Minimax sparse logistic regression for very high-dimensional feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1609–1622, Oct. 2013.
[29] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
[30] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.
[31] K. Kira and L. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Workshop Mach. Learn.*, 1992, pp. 249–256.
[32] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learn.*, 1994, pp. 171–182.
[33] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, 2003.
[34] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.
[35] T. G. Dietterich, "Machine learning research: Four current directions," *AI Mag.*, vol. 18, no. 4, pp. 97–136, 1997.
[36] K. Crammer, R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin analysis of the LVQ algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2002, pp. 462–469.
[37] Y. Li and B.-L. Lu, "Feature selection based on loss margin of nearest neighbor classification," *Pattern Recognit.*, vol. 42, no. 9, pp. 1914–1921, 2009.
[38] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.
[39] C. C. Chang and C. J. Lin. (2002). *Libsvm: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/papers/libsvm.ps.gz
[40] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *Ann. Statist.*, vol. 36, no. 2, pp. 614–645, 2008.
[41] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.
[42] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," Dept. EECS., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 797, 2010.
[43] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates, Inc., 2009, pp. 1348–1356.

Yun Li (M'10) received the Ph.D. degree in computer science from Chongqing University, Chongqing, China.

He was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is a Professor with the College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China. He has published more than 30 refereed research papers. His current research interests include machine learning, data mining, and parallel computing.

Dr. Li was the Co-Publication Chair of the 18th International Conference on Neural Information Processing in 2011. His research is currently sponsored by the National Natural Science Foundation of China and the Natural Science Foundation of Jiangsu.

Jennie Si (F'08) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, USA.

She has been with the faculty of the Department of Electrical Engineering at Arizona State University, Phoenix, AZ, USA, since 1991. She has also made new efforts to build a capability for studying some fundamental neuroscience questions in regards to the frontal cortex. Her lab is now well equipped with important techniques, such as multichannel single unit extracellular recording using a behaving rat model in chronic physiological experiments. Her current research interests include dynamic optimization using learning and neural network approximation approaches, namely approximate dynamic programming.

Dr. Si was an Associate Editor of the IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and the IEEE TRANSACTIONS ON NEURAL NETWORKS. She has served on several professional organizations' executive boards and international conference committees. She was the Vice President of Education with the IEEE Computation Intelligence Society from 2009 to 2012, an Advisor to the NSF Social Behavioral and Economical Directory, and served on several proposal review panels. She consulted for Intel, Arizona Public Service, and Medtronic. She is a Distinguished Lecturer of the IEEE Computational Intelligence Society, and an Action Editor of *Neural Networks*. She was a recipient of the National Science Foundation/White House Presidential Faculty Fellow Award in 1995, and the Motorola Engineering Excellence Award in 1995.

Guojing Zhou received the bachelor's degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, where he is currently pursuing the master's degree in computer science.

His current research interests include machine learning.

Shasha Huang received the bachelor's degree from Xuhai College, University of Mining and Technology, Xuzhou, China. She is currently pursuing the master's degree in computer science with the Nanjing University of Posts and Telecommunications, Nanjing, China.

Her current research interests include machine learning.

Songcan Chen received the B.S. degree in mathematics from Zhejiang University, Hangzhou, China, in 1983, the M.S. degree in computer applications from Shanghai Jiao Tong University, Shanghai, China, in 1985, and the Ph.D. degree in communication and information systems from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1997.

He has been a full-time Professor with the Department of Computer Science and Engineering at NUAA since 1998. He has authored and co-authored over 170 scientific peer-reviewed papers. His current research interests include pattern recognition, machine learning, and neural computing.